Cambridge Forum on AI: Law and Governance (2025), **1**, e46, 1–11
doi:10.1017/cfl.2025.10036

Cambridge Forum

RESEARCH ARTICLE

AI governance for military decision-making: A proposal for managing complexity

Osonde Osoba

Engineering and Applied Sciences, RAND Corporation, Santa Monica, CA, USA
Email: oosoba@rand.org

(Received 31 January 2025; revised 23 June 2025; accepted 13 October 2025)

Abstract

Military decision-making institutions face new challenges and opportunities from increasing artificial intelligence (AI) integration. Military AI adoption is incentivized by competitive pressures and expanding national security needs; thus, we can expect increased complexity due to AI proliferation. Governing this complexity is urgent but lacks clear precedents. This discussion critically re-examines key concerns that AI integration into resort-to-force decision-making organizations introduces. Beside concerns, this article draws attention to new, positive affordances that AI proliferation may introduce. I then propose a minimal AI governance standard framework, adapting private sector insights to the defence context. I argue that adopting AI governance standards (e.g., based on this framework) can foster an organizational culture of accountability, combining technical know-how with the cultivated judgment needed to navigate contested governance concepts. Finally, I hypothesize some strategic implications of the adoption of AI governance programmes by military institutions.

Keywords: AI governance; military decision-making; accountability; complex systems; deterrence

1. Introduction: charting a path from description to control[1]

The prospect of artificial intelligence (AI) in national security decision-making institutions raises moral and operational complications. I have highlighted key aspects of these complications in an earlier discussion (Osoba, 2024). My primary assertion was that any substantive decision-making institution (civilian or military) is a ***complex adaptive system*** because it features intelligent agents interacting and adapting under the influence of incentives and organizational structures. This article builds on that descriptive project to propose a minimal governance schema or framework for enabling better societal *control* of such complex institutions. The stability of the international political ecosystem requires responsible state actors to subject their national security operations (including decisions to resort to war) to international norms and ethical mandates. Effective decision system *control* is imperative to guarantee and to certify compliance with such normative standards.

[1]This is one of fourteen articles published as part of the *Cambridge Forum on AI: Law and Governance* Special Issue, *AI and the Decision to Go to War*, guest edited by Toni Erskine and Steven E. Miller. I would like to gratefully acknowledge Tuukka Kaikkonen for his careful review of this work, Mitja Sienknecht for stimulating discussions on accountability, and Luba Zatsepina for her helpful discussions on deterrence theory.

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/cfl.2025.10036 Published online by Cambridge University Press

AI already features at various levels of states' decision-making on whether and when to wage war. As an example, see the United States (US) development of automated tools to support more responsive processing for the large glut of intelligence, surveillance, and reconnaissance (ISR) data as reported by Harper (2018). ISR is a particularly relevant mission cluster because ISR data flows inform perceptions of the adversary's intent as well as deliberations on how and/or when to resort to force (jus ad bellum considerations). AI use is likely to proliferate in warfighting institutions in roles where these artefacts can be scoped to have a competitive advantage; the set of such roles is not an empty set, ISR and military logistics being examples (Economist, 2024).

The governance standard proposed in this article is simple. It emphasizes the importance of assuring the **trustworthiness/accountability** of the AI-deploying institution (a top–down *organizational* evaluation) as well as assuring the **robustness** of AI artefacts that contribute to the institution's missions or decisions (a bottom–up *technical* evaluation). As an institution's complexity grows, warrants of trustworthiness and accountability become especially crucial for maintaining its legitimacy and effectiveness. These evaluative dimensions (organizational and technical) are always important for taming complexity in mission-oriented institutions.

I argue, however, that the most important element of the governance standard is what I call the **culture of accountability** that a diligent implementation of any AI governance programme fosters. This culture is particularly important for an institution navigating the novel shock of AI proliferation. A culture of accountability in AI-extended organizations is evidenced by two key markers: First, the presence of deep technical expertise to rigorously validate that AI systems are fit for their intended purposes and operating environments. Second, the presence of strong faculties for ethical and normative deliberation needed to critically examine how the integration of AI artefacts affects the institution's value commitments.

The rest of this article goes as follows. In the next section, I present a sharpened description of key factors defining the new AI–human hybrid military decision-making organization. The factors discussed in that section identify new problems as well as new affordances introduced into resort-to-force decision-making institutions that adopt AI and automation. After this sharpened description, I turn to the question of how to control or govern such new hybrid organizations in section 3. I argue for a minimal governance schema for AI-augmented decision-making organizations. This schema or framework differs from other comprehensive AI governance frameworks in that I aim for a minimalist "value-agnostic" framework that can be targeted towards institution-specific norms. The idea is that any *jus ad bellum, jus in bello* or any other norms (including even *jus contra bellum* norms) can be built atop and enabled by such a framework. Section 4 examines some strategic implications of adopting an AI governance programme that complies with the proposed framework. I am particularly focused on the problem of deterrence for AI-hybrid organizations since deterrence calculi are central to resort-to-force decision-making among nation-states. Section 5 presents some concluding remarks to highlight the key themes from this article.

2. Understanding the new AI–human hybrid organization

The introduction of AI and autonomous systems in various roles in an organization's decision ecosystem will modify the institution's behaviours depending on the mode of AI/automation deployment.[2] We can expect further separation between the organization and the individuals who act in it, leading

[2]I use the terms "AI," autonomous systems, automated tools throughout the article. There are debates about the proper demarcations among these concepts. I am more focused on the fact of *constructed, non-human* forms of intelligence and agency than on the distinctions within that non-human set. I will tend to use the term "AI artefact" or "artificially intelligent artefact" to refer to these *constructed, non-human* forms of intelligence and agency. I make no distinctions based on the level or complexity of intelligence embedded in these artefacts. I do not think the systemic accountability crisis we are discussing here is strongly determined by the *level* of intelligence in these artefacts. It is their exponential proliferation and our deepening reliance that triggers the crisis.

https://doi.org/10.1017/cfl.2025.10036 Published online by Cambridge University Press

to an increased alienation of individuals from their actions (Han, 2018, p. 13). This renders the character of AI-augmented decision-making institutions less legible without new frameworks for systems comprehension and management. For comparison, consider the evolution in our collective understanding of economic production systems before and after the introduction of industrial-age machines. There was significant uncertainty about the consequences of industrialization while the process of industrialization was starting up. Or, to use a less common example, consider the evolution in our collective understanding of state governance under the influence of comprehensive quantification and statistical practices introduced in the 1800s (especially in post-revolution France; Porter, 1996). A 17th century state official would have difficulty understanding the limits of what is quantifiable or knowable about a modern nation's economy or public health without some framework for understanding modern statistical quantification.

We can try to identify some foundational factors in any framework for comprehending our new AI-augmented decision-making institutions. The first factor is the concept of *expanded* ***accountability*** *in human–machine hybrid systems*. I agree with other scholars who point out that accountability requires the ability to be held *responsible* for outcomes.[3] Floridi (2016, p. 6) scopes an agent's condition of being responsible to mean being "causally accountable for a state […] and, therefore, as a consequence, of being morally answerable (blameable/praisable) for its state." *Responsibility* includes the capacity to pursue explicit goals and (crucially) to meaningfully bear blame for adverse outcomes. The capacity to meaningfully bear blame requires a capacity for redress or remedy and even, potentially, to bear punishment. Floridi (2016, p. 6) refers to this as being able to "learn from, and modify" one's own behaviour. We may also include the capacity to be *forgiven* as a twin to the capacity to bear punishment, as Hannah Arendt (Arendt, 1958, pp. 236–43) does. In Arendt's account, both capacities (bearing blame/praise and being forgiven) can be viewed as attempts to close out or *settle* irreversible actions that have gone awry, as is bound to happen since we are always subject to unpredictability even under the most favourable conditions and the best of intents. The capacity to engage in such *settling moves* is necessary for humans (or, more generally, social agents, including AI agents) to socialize and do politics in an uncertain world.

AI and automation artefacts *do not* meet these requirements. Such artefacts fail to meet even limited conceptions of responsible personhood by current legal (and even looser cultural) standards (Osoba, 2024). This leaves us with the problem of trying to deploy AI artefacts that are (partially) autonomous in an accountable manner without the ability to hold these artefacts accountable in any meaningful way. Potential resolutions to this conundrum include Sienknecht's (2024) concept of "proxy responsibility," Floridi's (2016) concept of "distributed moral responsibility," and other approaches that would allow an artefact's trustworthiness to be tangibly rooted in a network of responsible non-artificial agents deploying the artefact. Accountability is important for anchoring the trustworthiness of decision-making institutions.

The second factor that can help us understand AI-augmented decision-making institutions is ***cognitive diversity*** *in hybrid systems* that feature humans as well as AI and automation artefacts. Here, cognitive diversity refers to complementarity in skill and task competencies between human and

[3]What follows in the text is a mainly consequentialist account of responsibility, in the sense that it only focuses on actions and their downstream effects and responses. Some accounts of responsibility make room for reasoning about the agent's intentions, e.g., in Strasser's account of "full-fledged moral agency" (Strasser, 2022, pp. 525–7). I do not find intentions useful for scoping the concept of responsibility for this discussion because we do not have a stable concept of what it means for artificial agents to *intend*. I will stick to consequentialism because I read other cogent accounts that have sought to define intention in terms of actions as reducible to forms of consequentialism. For example, in Anscombe's account of intention, she says that "the primitive sign of wanting is trying to get" (Anscombe, 1957, p. 68). In other words, we recognize the form of an intention ("wanting") by the actions it produces in the agent (the observable attempt at getting something). Anscombe's conception of intention may evidently be applied to any agent that produces effects or actions in the world, including artificial agents. However, it roots intentions in the observable traces of those effects or actions. I read that position as essentially reducible to consequentialism.

https://doi.org/10.1017/cfl.2025.10036 Published online by Cambridge University Press

AI/automation agents.[4] AI systems have the relative advantage in their ability to scale up actions. Human agents (so far) have the relative advantage in their ability to account for nuanced contextual clues as well as in strategic deliberation under conditions of incomplete or ambiguous information. As another example, in the context of visual recognition tasks, AI systems and human agents are sensitive to different kinds of deceptions in images. Cognitive diversity is a design consideration for AI-deploying institutions, not an inevitable disadvantage or advantage. Institutions can make (un)wise use of cognitive diversity in their organizational structure to produce (in)effective decision-making processes.

The third factor is the joint concept of ***human deskilling and specialization*** in human–machine hybrids. Task specialization and task deskilling may be viewed as two sides of the same coin. Widespread and competitive specialization of AI artefacts to specific task sets creates marginal pressure for humans to specialize in complementary tasks which leads to human collective deskilling in the AI-targeted task set.[5]

This deskilling dynamic is often discussed as a negative externality to be avoided. I argue, by analogy to the division of labour in industrial economies (another complex adaptive system), that this negative view of deskilling is, however, an incomplete account.

We can view the deskilling concept from a different angle with the help of a (heavily) simplified application of Riccardo's concept of *comparative advantage* (Costinot & Donaldson, 2012) to the implications of cognitive diversity. Take the hypothetical example of two different agents producing goods to exchange for money. Suppose each agent has different, possibly complementary competencies. Under some assumptions, the theory of comparative advantage points out that (1) the joint "economy" of the two agents is more productive when each agent specializes in producing the goods for which they have greater relative competence and then trade surpluses with each other; and (2) each agent also individually reaps better profits at the end of such exchanges.

Cognitive diversity in a human–machine hybrid team is a statement about differences in relative competences. In a group of agents of different relative competences, tasks can often be reallocated to improve the combined group's efficiency. For example, suppose we judge the total productivity of a decision-making institution by the quantity of relevant good decisions it can inform or make, relative to the amount of attention (human or AI) applied. We can identify the value that accrues to individual agents as the sum of their partial contributions to the total set of decisions given a fixed amount of attention applied. Under certain assumptions, a comparative advantage analysis suggests that the overall performance of an AI-deploying decision-making institution can be improved via task specialization. This is true even when specialization is accompanied by specialization's counterpart, deskilling.

The foregoing discussion assumes that task roles are somewhat interchangeable between humans and intelligent artefacts even if there are relative competencies. This interchangeability makes the prospect of efficient task reallocation between humans and machines potentially viable.

A caveat here: the capacity for moral deliberation counts as a relative competence and domain of comparative advantage for the human agents. This is relevant especially when the requisite form of moral deliberation is framed as more than a mere quantitative optimization of expected "utilities." For example, moral deliberation may be based on participatory processes that are focused more on the social act of honouring the voices of impacted stakeholders. Allocating moral deliberation tasks to machines in a human–machine hybrid team in such a setting would be inefficient. This raises a limiting case for consideration: if the human agent's entire contribution to a decision-making process

[4]I am not aware of any prior use of this term to signify this specific concept in the literature. The closest work is Hernández-Orallo and Vold's (2019) detailed characterizations of cognitively-extended humans. They discuss cases in which humans and machines can be "synergetic" (p. 509) and catalogue a non-exhaustive set of modes of AI cognitive extension (pp. 510–11). However, they do not use the term "cognitive diversity" to describe synergetic applications of AI and humans to a workflow.

[5]This premise hides a key factor that drives this human deskilling cycle: capital's revealed preference for replacing human labour where it is possible to do so.

https://doi.org/10.1017/cfl.2025.10036 Published online by Cambridge University Press

is moral deliberation and/or the capacity to bear moral responsibility, then no further reallocation of tasks to artefacts will improve decision-making.

This limiting case raises questions about the quality of moral deliberation when humans rely on AI-based cognitive extensions. Cummings (2017) documents degradation in operational skills when human operators over-rely on AI tools in laboratory settings (automation bias). Schwarz (2020) suggests that this observed operational skill degradation also extends to moral deliberation. I argue that these findings are actually observations of the effects of inefficient task allocation (including robust moral deliberation tasks) in human–machine teams.

These factors paint a picture of new AI-hybrid organizations with more efficient task allocation and potentially fewer decision burdens on human actors. However, in such organizations, it is also harder to cleanly ascribe responsibility for decision outcomes. This results in a potentially more effective organization but with weaker lines of accountability.

These factors help us better understand our new AI–human hybrid decision-making institutions. But how do we better govern them?

3. Governing complex hybrid decision-making institutions

Military decision-making institutions are poised to become even more complex as they bear the shock of AI integration. The ultimate effect of this integration will hinge critically on the capacity of these institutions to make wise *responsible* choices in deploying and governing AI. We aim to establish a foothold here by addressing the following question:

*How might we responsibly **govern** the use of AI in military decision-making organizations?*

This discussion will raise at least the following implications for reflection: do AI governance efforts in these more complex hybrid institutions impose a strategic advantage or disadvantage (*strategic latency*; Davis & Nacht, 2017)? Can AI governance improve the accountability of military decision-making organizations to civil society in liberal-democratic states? I address these strategic questions in the next section after outlining my proposal for AI governance.

3.1 An Is-ought distinction

I distinguish this discussion's target question from the related question of "***should** AI be used in military contexts at all?*" These questions of "ought" or "should" around military AI use are moot, or at least, not timely for two reasons: (1) AI deployment in the military is highly incentivized by both nation-state competition and the increasing complexity of the national security environment (Osoba, 2024); and (2) AI is already in use in some military contexts (The Economist, 2024). By sidestepping the "ought" question, I am not conceding that a military organization's choice to adopt AI is beyond reproach or free of "tragic" implications (Renic, 2024). For a prime example of such a tragic implication, consider the argument that the use of AI and automation to mediate violent acts may further deaden the emotional impact of violent acts (Renic, 2024) and erode norms of restraints in war (Erskine, 2024). However, if we concede that there are strong structural factors that privilege AI accelerationism within military institutions, the practical duty is to reflect on how to govern AI's responsible use. This approach has more potential for guiding organizations' actions towards normative moral ends.

3.2 AI governance: standards and mechanisms over norms

Governance is concerned with organizational monitoring and control in service of legal or moral ends. I focus here on highlighting AI governance infrastructures that scope and enable governance. I do this without specifying target moral norms for AI governance. Moral norms as implemented via governance infrastructures will differ among military institutions. An AI governance infrastructure can

https://doi.org/10.1017/cfl.2025.10036 Published online by Cambridge University Press

be flexible enough to be repurposed into the service of differing moral standards. On the other hand, the relevant moral norms are deeply contingent on the institution. We can think of this approach as a study of frameworks and tools for cultivating virtues instead of a study of a specific virtue being cultivated. Or we can think of this approach as similar to a study of voting mechanisms instead of studying the various kinds of political structures that voting mechanisms can support. The goal is to support any emergent or self-organized norms (Winston, 2023) that may evolve from this complex adaptive system, not to dictate the norms themselves. In fact, a shared AI governance infrastructure can be used by stakeholders to catalyse the self-organization of more value-laden norms. In that sense, we may think of this project as an exercise in setting an anticipatory norm (Prem, 2022) to enable further governance.

This focus on mechanisms makes for a more fertile discussion because an institution's target moral norms are contingent and often in flux. As a case in point, review the remarkable variation in what commercial institutions mean by "responsible" AI use (Biden, 2023; de Laat, 2021; Khan et al., 2022; National Institute of Standards and Technology, 2023). The norms of what is considered "responsible" AI use varies wildly across institutions and includes standards like accountability, fairness, transparency, privacy, equity, non-discrimination, civil rights, reliability, robustness, safety, security, etc. This kind of normative fragmentation hampers the portability of governance mechanisms across institutions. The strategy here is instead to focus on standards that are shared across mission-oriented institutions and anchor a governance framework on those.

3.3 Proposed minimal standards for AI governance in military decision-making

There are only two elements in my minimal proposal for an AI governance infrastructure. They align with a decision-making institution's need to:

(a) use verifiably reliable tools to achieve its ends; and
(b) provide warrants of appropriate behaviours and modify their behaviour when pressured.

These governance standards are broadly applicable and relatively agnostic to the specific moral norms we would impose on a decision-making institution. In more detail, the two elements may be summarized as the following two cluster concepts.

3.3.1. Reliability/robustness

This cluster concept is concerned with assuring that deployed AI artefacts are reliable and fit for purpose across a wide range of expected and unexpected operating scenarios. Reliability and robustness are mainly properties of individual artefacts and subsystems. They activate technical or scientific modes of evaluation (*techne* or *episteme*). They may be operationalized as technical measurements of properties of the AI artefact under scrutiny. Measurement constructs under this umbrella include concepts like accuracy, safety, resilience, stability, reliability, etc. In the military and adversarial context, this concept needs to be broader than just "reliability." AI and automation artefacts in military contexts must withstand a constant onslaught of deception and manipulation. To illustrate this point, consider civilian use of automated cars and drones. The standard operating environments for these devices are truly complex, as the myriad start-ups in the autonomous vehicle space attest. But however complex these civilian operating environments are, they pale in comparison to the complexity of autonomous land or air operations in hostile or contested environments.

3.3.2. Trustworthiness and accountability

This cluster concept captures both an AI-extended institution's capacity to be bound by regulations and other constraining norms and the institution's ability to provide truthful evidence of that capacity. Relevant constraining norms can include the laws of war (*jus ad bellum*, in particular, in relation to resort-to-force decision-making), specific modes of transparency, as well as modes of human

https://doi.org/10.1017/cfl.2025.10036 Published online by Cambridge University Press

oversight. An institution's trustworthiness and accountability are precisely the factors that signal an institution's ability to bear responsibility and its capacity to learn from and modify its behaviour in response to feedback (for example, when found to be in violation of norms). Recall that I argued that these capacities are requirements for agentic accountability. I am now applying these to requirements to AI-extended institutions. Trustworthiness and accountability are also hierarchical concepts in the sense that a larger system/institution's trustworthiness is supported by the ability to meaningfully account for the behaviour of it sub-parts.

This concept has a structural flavour: It is not sufficient to provide evidence for an institution's or AI system's norms compliance at a specific moment in time. To satisfy trustworthiness and accountability, it is more important to give evidence of structural and procedural measures that monitor and assure the institution's norms compliance over time. As an example, it is *necessary but insufficient* evidence of trustworthiness to observe that cars produced by an institution seem to operate safely. It is more important to *transparently certify* that the carmaker has up-to-date internal processes for verifying that the cars it produces are safe.

In this carmaker example, we see that transparency (giving evidence and certifying) is necessary for trustworthiness and accountability. I do not elevate transparency to a first-class element of the framework because trustworthiness is the ultimate purpose (a *final cause* in the Aristotelian sense) of transparency practices. Transparency is rarely an end-in-itself.

Trustworthiness implicates reliability but not necessarily vice versa.[6] It is an organizational property since it is a function of processes and accountable human agents subject to internal and external incentives. The mode of evaluation it activates involves more practical wisdom and highly contextual local organizational knowledge or *mētis*, to use James Scott's (2020) term.

Hereafter, I will use the term *AI governance* to refer to any coherent set of practices and programmes that aim to implement at least this minimal set of AI governance standards. The exact form of such implementations will necessarily vary.

3.4 AI governance for military institutions?

Governance processes are burdensome and effortful especially for mission-oriented institutions. Without a clear link to mission effectiveness or readiness, the work of governance begins to look like mere theatre. This perception is part of my reason for positing a minimal/parsimonious governance standard that can be anchored in norms while having clear strategic utility. Arguing for the strategic utility of a more capacious governance (or ethical) standard requires more careful justification. One argument for more capacious ethical standards in war is that adversaries that adhere more strictly to ethical conduct in war may have an easier time collaborating during post-war repair. Adversaries that do not resort to extensive atrocities during times of war may have an easier time repairing relations once hostilities cease. This continues to be a good argument even if recent experience show that the argument does not deter all atrocities.

I argue instead for the strategic utility of a minimal governance standard outlined above. The heart of the argument is that AI governance contributes to mission effectiveness and, more importantly, to a ***culture of accountability***. The argument does not require perfect or effective implementation of the governance standard. A mere diligent pursuit of the standard may be sufficient to reap some utility from it.[7]

[6]In some sense, the cluster concept of trustworthiness and accountability cashes out as technical reliability at the level of an individual artificial agent.

[7]Lindblom's (1979) old concept of "muddling through" recurs in literature on control in complex systems, e.g., in Scott (2020) and in Davis, McDonald, Pendleton-Jullian, O'Mahony and Osoba (2021). The goal is to intervene a policy system. Actors formulate and execute an initial imperfect plan of attack under the assumption that further course corrections will be needed to achieve the desired outcome eventually. Iris Murdoch's discussion of "The Sovereignty of Good" (Murdoch, 1970) provides further support for this strategy of a diligent (even if imperfect) pursuit of a desirable imprecise standard.

https://doi.org/10.1017/cfl.2025.10036 Published online by Cambridge University Press

Military decision-making includes many high-risk use cases as well as natively adversarial scenarios. We anticipate that there will be extensive attempts to deceive and to thwart actions made through these pipelines. Decided through these pipelines, Geist (2023) paints a more catastrophic version of this observation in the context of nuclear deterrence. He argues that the proliferation of AI in warfare may lead to "a deception-dominant world in which states are no longer able to estimate their relative capabilities." Effective deployment of AI will need to be *robust* to such deception-laden scenarios. Frontloading a commitment to robustness reduces the likelihood of catastrophic failures that naïve AI integration can cause.

A basic implication of the governance standard for accountable AI use is the maintenance of constant situational awareness of where and for what purposes AI artefacts are deployed within organizations. This aspect of AI governance has been quite challenging for private sector stewards of "AI First" platforms.[8] One of the main pain points in recent efforts to comply with the EU's DMA (European Parliament, 2022) and AI Act (European Commission, 2021) has been the need for platforms to maintain detailed comprehensive inventories of AI and data systems. Such inventories are important for tracking and governing prescribed risks; for example, in the case of the DMA, tracking the risk of non-privacy-compliant uses of specific data sources. Frontloading a commitment to accountability in the use of AI would incentivize careful cataloguing of AI deployments, giving military decision-making institutions an early advantage in governing their use of AI.

The most important point concerns the *organizational culture* that the adoption of an AI governance standard fosters. The full scope of good AI governance is still subject to contestation (see earlier point about variation in conceptions of Responsible AI use). Full resolution is not likely in the near term given the pace of major innovations in AI. The sharp difference in the governance needs for classical vs. generative AI models (like ChatGPT) illustrates the field's near-constant instability.

Given this uncertain governance regime, we may find it useful to glean insights from governance innovations during the rise of a different technology: that of quantification and standardization (Porter, 1996, pp. 49–51). In his account of the rise of quantification and standardization in state governance, Porter (1996) describes a culture of technocratic governance ("Technocra[cy] in the French tradition," p. 146) that combines a deep quantification facility with *cultivated* expert *judgment* capable of flexibly balancing social and moral constraints when serving the needs of their constituents. In this tradition, it is insufficient for the practitioner to simply appeal to technical quantifications or engineering measurements (*techne*) to justify choices and decisions. The practitioner must consider the operating context, the plural perspectives of relevant local stakeholders, the operating ethical mandates, and a finely calibrated understanding of potential downstream effects of various technical interventions. And, most crucially, the practitioner must apply practical wisdom and cultivated judgment (*mētis*) to balance all these considerations when settling on a decision. This culture of balanced faculties for governance is precisely what is required to manage complex ecosystems of AI-equipped decision pipelines while both negotiating and giving a clear account of the value-laden standards that have been adopted. This culture of balanced governance faculties is what I refer to as a culture of accountability in AI-extended organizations. Without this culture, navigating contested AI governance concepts is hard.

A cultivated judgment faculty for balancing normative constraints is especially useful in liberal democratic polities where the military is *accountable* to civil society for its conduct.[9] The cultivated faculty enables military decision-making institutions to publicly perform their moral deliberations about AI deployment in military contexts in a credible way. Such public performances of moral

[8] I.e. multinational digital-native firms like Google, Meta, Microsoft, etc.

[9] Military accountability to civil society is often an indirect, intermediated process. For example, an effective military will always need to hide some operational secrets from the general population. Trustworthiness and accountability only become more important when secrecy is necessary. This is often achieved by clearly communicating key norms and giving evidence of adherence to those norms.

https://doi.org/10.1017/cfl.2025.10036 Published online by Cambridge University Press

deliberation can bolster the military's legitimacy. Jumpstarting AI governance programmes in military decision-making institutions, specifically programme elements targeting *trustworthiness and accountability* in AI extension, can cultivate institutional capacity for both the quantification and the moral deliberation faculties needed to govern large AI-equipped decision-making institutions.

4. Strategic implications of AI governance

Suppose a nation's military decision-making institutions choose to adopt AI governance programmes. Are there strategic implications to that choice? The first immediate concern is about deterrence.

How would the adoption of AI governance practices in military decision-making organizations affect deterrence calculi, if at all? Answering this question requires some speculation. But let us attempt an informed speculation starting from first principles. The aim of deterrence is to restrain a security actor (the aggressor), often by means of threats, from taking unwanted action (e.g., attacking oneself or an ally; Mazarr, 2018). Successful deterrence is sometimes thought to require the activation of two subjective perceptions in the mind of the aggressor: the deterring state's sufficient *capability* to defend and the deterring state's *will* to carry out the implied threat.

The adoption of an AI governance programme can become relevant to deterrence calculi *if* the governance procedures retard the deterring state's ability to make timely observations and resort-to-force decisions. This is plausible if, for example, the governance processes impose additional levels of confirmatory review to ISR workflows. This can also happen if lines of accountability are so muddied that detection alarms slip through the reporting cracks. A resourceful and attentive aggressor may then be willing to gamble on the chance of achieving total victory within the decision delay window caused by slowed governance. There is also the possibility that, since good AI governance may increase a military decision-making institution's transparency, it may render the institution more vulnerable to espionage.

Less negative outcomes are possible. For example, an aggressor adopting effective AI governance may become more operationally effective at its military goals relative to the deterring state. This can happen if better governance improves the institution's ability to carefully deploy reliable and effective AI systems in a mission-targeted manner. There is also no obvious connection between the aggressor's perception of the deterring state's capability and the deterring state's adoption of AI governance in its military decision-making organizations. The same applies to perceptions of the deterring state's will to act.

A second strategic concern is about the dual-use implications or what Davis and Nacht (2017) refer to as "strategic latencies" in technologies. AI governance programmes may be viewed as organizational and social technologies that aim to enable better normative control and mission effectiveness for human–AI decision-making hybrids. The most obvious negative effect of a purely operational instantiation of AI governance is that it makes unscrupulous aggressor states more effective in their applications of AI to military decision-making. The obvious positive effect is that it makes the use of AI more responsive to a conscientious security actor's moral standards. It can also have the effect of making the moral commitments of a military decision-making institution *more legible* to external observers.

5. Concluding remarks

The aim of this piece has been to highlight frameworks and concepts that may help to manage the increasing complexity of military decision-making organizations as they navigate the shock of proliferating AI. My primary suggestion for better control is the adoption of a parsimonious responsible AI governance framework. I contend that adopting and implementing such an AI governance framework addresses many operational goals of decision-making institutions and helps to foster a *culture of accountability* that is instrumental for both operational and ethical ends. In speculating about the

https://doi.org/10.1017/cfl.2025.10036 Published online by Cambridge University Press

potential strategic implications of AI adoption, I suggest some negative implications for deterrence calculi, primarily under the scenario that the adoption of AI governance processes results in slower decision-making.

A final note about my supposedly "value-agnostic" approach. There are no truly value-agnostic mechanisms. There is no value-free normative vantage point from which we can lever ourselves into a world of either fully compliant wars or no wars whatsoever. Likewise, the governance standards discussed above are also not value-free; they embed a commitment to transparent and responsive governments (e.g., a functioning liberal democracy). Without that minimum bar, the whole discussion is moot.

Funding Statement. None to declare.

Competing Interests. None to declare.

References

Anscombe, G. E. M. (1957). *Intention*. Blackwell.

Arendt, H. (1958). *The human condition* (2nd ed.) University of Chicago Press.

Biden, J. R. (2023). Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. The White House. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence

Costinot, A., & Donaldson, D. (2012). Ricardo's theory of comparative advantage: Old idea, new evidence. *American Economic Review*, *102*(3), 453–458.

Cummings, M. L. (2017). Automation bias in intelligent time critical decision support systems. In *Decision making in aviation* (pp. 289–294). Harris Don, Routledge.

Davis, P. K., McDonald, T., Pendleton-Jullian, A., O'Mahony, A., & Osoba, O. (2021). Reforming the teaching and conducting of policy studies to deal better with complex systems. *Journal on Policy and Complex Systems*, *7*(1), 1–15.

Davis, Z., & Nacht, M. (2017). Strategic Latency red, white, and blue: Managing the national and international security consequences of disruptive technologies. (Report No. LLNL-BOOK-746803). Lawrence Livermore National Lab.

de Laat, P. B. (2021). Companies committed to responsible AI: From principles towards implementation and regulation? *Philosophy and Technology*, *34*(4), 1135–1193.

Economist. (2024, June 20). How AI is changing warfare. https://www.economist.com/briefing/2024/06/20/how-ai-is-changing-warfare

Erskine, T. (2024). *Before* Algorithmic Armageddon: Anticipating immediate risks to restraint when AI infiltrates decisions to wage war. *Australian Journal of International Affairs*, *78*(2), 175–190. https://doi.org/10.1080/10357718.2024.2349598

European Commission. (2021). *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*. COM/2021/206 final. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

European Parliament. (2022). *Regulation (EU) 2022/1925 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act)*. https://eur-lex.europa.eu/eli/reg/2022/1925/oj

Floridi, L. (2016). Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2083), 20160112.

Geist, E. (2023). *Deterrence under uncertainty: Artificial intelligence and nuclear warfare*. Oxford University Press.

Han, B. C. (2018). *What is Power?* Polity.

Harper, J. (2018). Artificial Intelligence to Sort through ISR Data Glut. *National Defense*, *102*(770), 33–35.

Hernández-Orallo, J., & Vold, K. (2019). AI extenders: The ethical and societal implications of humans cognitively extended by AI. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 507–513.

Khan, A. A., Badshah, S., Liang, P., Waseem, M., Khan, B., Ahmad, A., Fahmideh, M., Niazi, M., & Akbar, M. A. (2022). Ethics of AI: A systematic literature review of principles and challenges. *Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering*, 383–392.

Lindblom, C. E. (1979). Still muddling, not yet through. *Public Administration Review*, *39*(6), 517–526.

Mazarr, M. J. (2018). *Understanding deterrence*. RAND Corporation. https://www.rand.org/pubs/perspectives/PE295.html

Murdoch, I. (1970). *The Sovereignty of Good*. London: Routledge.

National Institute of Standards and Technology. (2023, January 26). *Artificial intelligence risk management framework (AI RMF 1.0)*. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

Osoba, O. A. (2024). A complex-systems view on military decision making. *Australian Journal of International Affairs*, *78*(2), 237–246. https://doi.org/10.1080/10357718.2024.2333817

https://doi.org/10.1017/cfl.2025.10036 Published online by Cambridge University Press

Porter, T. M. (1996). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press.

Prem, B. (2022). Governing through anticipatory norms: How UNIDIR constructs knowledge about autonomous weapons systems. *Global Society*, *36*(2), 261–280.

Renic, N. (2024). Tragic reflection, political wisdom, and the future of algorithmic war. *Australian Journal of International Affairs*, *78*(2), 247–256. https://doi.org/10.1080/10357718.2024.2328299

Schwarz, E. (2020). Silicon Valley goes to war: Artificial intelligence, weapons systems and the de-skilled moral agent. *Philosophy Today*, *64*(3), 485–505.

Scott, J. C. (2020). *Seeing like a state: How certain schemes to improve the human condition have failed*. Yale University Press.

Sienknecht, M. (2024). Proxy responsibility: Addressing responsibility gaps in human-machine decision making on the resort to force. *Australian Journal of International Affairs*, *78*(2), 191–199. https://doi.org/10.1080/10357718.2024.2327384

Strasser, A. (2022). Distributed responsibility in human–machine interactions. *AI and Ethics*, *2*(3), 523–532.

Winston, C. (2023). International norms as emergent properties of complex adaptive systems. *International Studies Quarterly*, *67*(3), sqad063.

Cite this article: Osoba, O. (2025). AI governance for military decision-making: A proposal for managing complexity. *Cambridge Forum on AI: Law and Governance*, *1*, e46, 1–11. https://doi.org/10.1017/cfl.2025.10036

https://doi.org/10.1017/cfl.2025.10036 Published online by Cambridge University Press